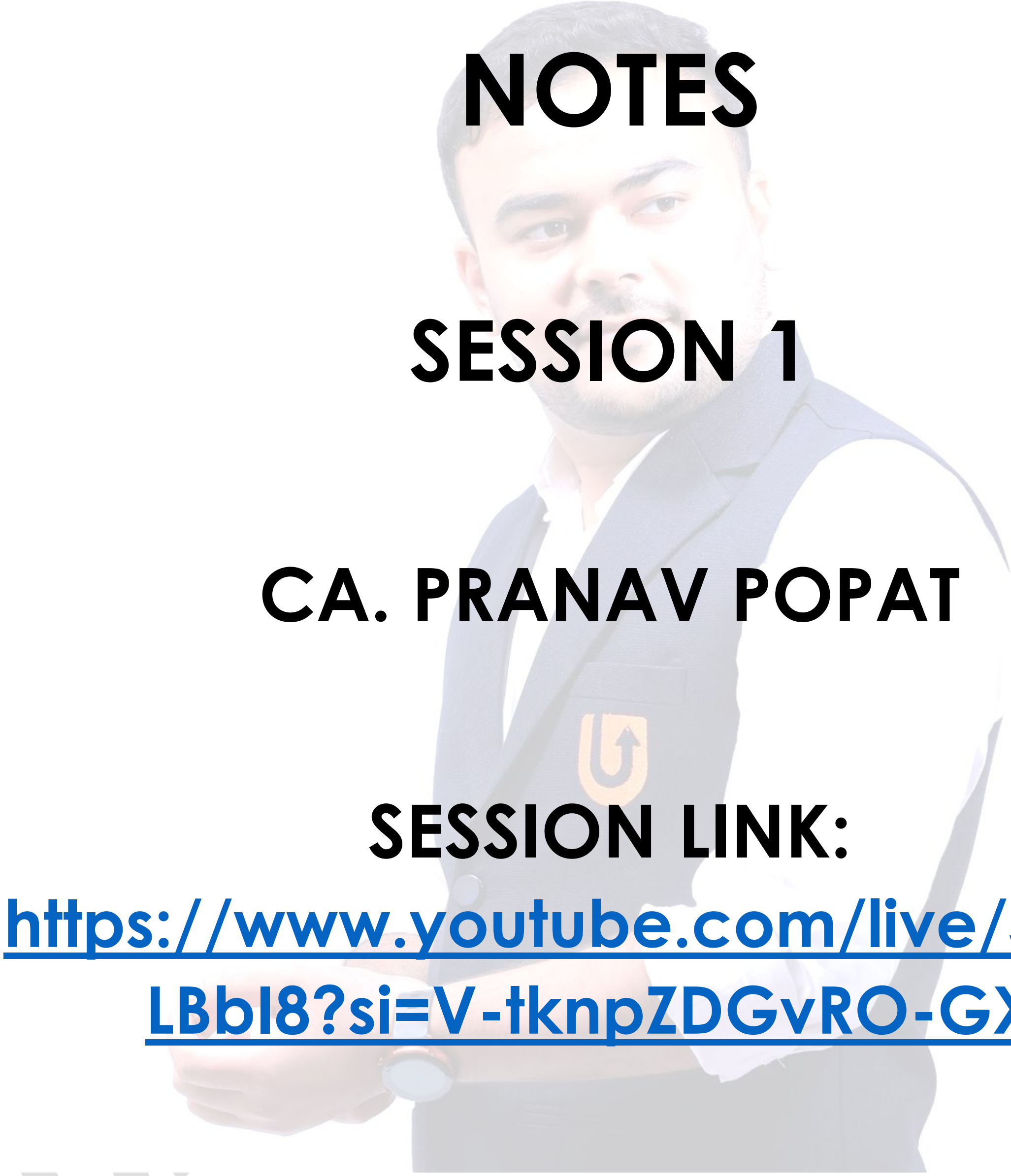
THEORY HAI ZAROORI NOTES



SESSION 1

CA. PRANAV POPAT

SESSION LINK:

https://www.youtube.com/live/58T5_vLBbI8?si=V-tknpZDGvRO-GX1

JOIN TELEGRAM CHANNEL FOR ALL UPDATES AND NOTES:

https://telegram.me/learnwithpranav

THEORY WEIGHTAGE

Chapter	13. Statistical Description of Data	14. Central Tendency & Dispersion	15. Probability	16. Theoretical Distribution	17. Correlation & Regression	18. Index Numbers	Total
May 18	2	4	2	3	6	8	**25**
Nov 18	6	1	0	0	2	3	**12**
Jun 19	5	3	1	0	1	5	**15**
Nov 19	1	7	0	2	2	5	**17**
Nov 20	8	5	0	4	3	6	**26**
Jan 21	10	5	1	2	2	4	**24**
Jul 21	6	1	0	0	1	0	**8**
Dec 21	3	5	0	0	2	4	**14**
Jun 22	9	3	0	1	4	6	**23**
Dec 22	4	3	1	2	1	3	**14**
Jun 23	2	0	0	0	0	2	**4**

THEORY CONCEPTS

Statistical Description of Data – Basics of Statistics

<table>
<tr><td><b>Definition of Statistics</b></td><td>
<ul>
<li>Plural Sense: Any data – quantitative or qualitative used for statistical analysis.</li>
<li>Singular Sense: Scientific method of collecting, analyzing, and presenting data to draw statistical inferences. conclusion It is also called as Science of Averages or Science of Counting</li>
</ul>
</td></tr>
<tr><td><b>Origin of Word</b></td><td>
<table>
<tr><th>Language</th><th>Actual Word</th><th>Memorize by</th></tr>
<tr><td>Latin</td><td>Status</td><td>Latus</td></tr>
<tr><td>Italian</td><td>Statista</td><td>Pasta</td></tr>
<tr><td>German</td><td>Statistic</td><td>Breadstick</td></tr>
<tr><td>French</td><td>Statistique</td><td>Barbeque</td></tr>
</table>
</td></tr>
<tr><td><b>Publication</b></td><td>
<table>
<tr><td><b>Koutilya’s Arthashastra</b></td><td><ul><li>Record of Birth and Deaths</li><li>Chandragupta’s reign</li><li>4th Century B.C</li></ul></td></tr>
<tr><td><b>Abu Fezal’s Ain-i-Akbari</b></td><td><ul><li>Record on Agriculture</li><li>Akbar Reign</li><li>16th Century A.D.</li></ul></td></tr>
<tr><td><b>First Census</b></td><td><ul><li>Egypt 300 BC to 2000 BC</li><li>By Pharaoh</li></ul></td></tr>
</table>
</td></tr>
<tr><td><b>Application of Statistics</b></td><td>
<ul>
<li>Economics: Demand Analysis, Future Projection etc.</li>
<li>Business Management: Decision making using quantitative techniques not intuition</li>
<li>Industry and Commerce: Profit maximization using business data – sales, purchase, market etc. by consulting experts</li>
</ul>
</td></tr>
<tr><td><b>Limitation of Statistics</b></td><td>
<ul>
<li>It deals with aggregate data and not individual data</li>
<li>Quantitative data can only be used, however for qualitative – it needs to be converted into quantitative</li>
<li>Projections are based on conditions/ assumptions and any change in that will change the projection. Example: Future projections of sales</li>
<li>Sampling based conclusions are used, improper sampling leads to improper results. Random Sampling is must.</li>
</ul>
</td></tr>
<tr><td><b>Data</b></td><td>
<ul>
<li>Quantitative Information shown as number</li>
<li>Primary: first time collected by agency/ investigator</li>
<li>Secondary: collected data used by different person/ agency</li>
</ul>
</td></tr>
</table>

<table>
<tr>
<td><b>Variable</b></td>
<td>
<ul><li>Measurable Data – Value can vary</li></ul>
<table>
<tr><td><b>Discrete Variable</b></td><td><ul><li>When a variable assumes a finite or countably infinite isolated values.</li><li>Example: no. of petals in a flower, no. of road accident in locality</li></ul></td></tr>
<tr><td><b>Continuous Variable</b></td><td><ul><li>When a variable assumes any value from the given interval (can also be in decimals, fractions).</li><li>Example: height, weight, sale, money</li></ul></td></tr>
</table>
</td>
</tr>
<tr>
<td><b>Attribute</b></td>
<td><ul><li>Qualitative Characteristics. Example: gender of a baby, the nationality of a person, the colour of a flower etc.</li></ul></td>
</tr>
<tr>
<td><b>Collection of Primary Data – Interview Method</b></td>
<td>
<table>
<tr><th>Method</th><th>Details</th></tr>
<tr><td><b>Personal Interview</b></td><td><ul><li>Where data is collected directly from respondents.</li><li>Highly Accurate – Low Coverage</li><li>Example: Natural Calamity, Door to Door Survey</li></ul></td></tr>
<tr><td><b>Indirect Interview</b></td><td><ul><li>When reaching respondent is difficult, data is collected by contacting associated persons.</li><li>Highly Accurate – Low Coverage</li><li>Example: Rail accident</li></ul></td></tr>
<tr><td><b>Telephone Interview</b></td><td><ul><li>Data is collected over phone</li><li>Quick and non-expensive method</li><li>Low Accuracy – High Coverage</li></ul></td></tr>
</table>
</td>
</tr>
<tr>
<td><b>Collection of Primary Data – Mailed Questionnaire Method</b></td>
<td><ul><li>In this method well drafted and soundly sequenced questionnaire,</li><li>covering all the important aspects of the data requirement is sent to respondent for filling.</li><li>Here coverage is wide but amount of non-responses will be maximum</li></ul></td>
</tr>
<tr>
<td><b>Collection of Primary Data – Observation Method</b></td>
<td><ul><li>In this method data is collected by direct observation or using instrument.</li><li>For example: data on height and weight for a group of students.</li><li>Although more accurate but it is time consuming, low coverage and laborious method.</li></ul></td>
</tr>
<tr>
<td><b>Collection of Primary Data – Questionnaire Filled and sent by Enumerators</b></td>
<td><ul><li>Mix of Interview and Mailed Questionnaire</li><li>Enumerator means a Person who directly interacts with respondent and fills the questionnaire.</li><li>It is generally used in case of Surveys and Census.</li></ul></td>
</tr>
</table>

<table>
<tr><td><b>Sources of Secondary Data</b></td><td>
<table>
<tr><td><b>International Sources</b></td><td>World Health Organization (WHO), International Monetary Fund (IMF), International Labor Organization (ILO), World Bank</td></tr>
<tr><td><b>Government Sources</b></td><td>In India – Central Statistics Office (CSO), Indian Agricultural Statistics by the Ministry of Food and Agri, National Sample Survey Office- NSSO, Regulators – RBI, SEBI, RERA, IRDA</td></tr>
<tr><td><b>Private or Quasi-govt. sources</b></td><td>Indian Statistical Institute (ISI), Indian Council of Agriculture, NCERT</td></tr>
</table>
</td></tr>
<tr><td><b>Scrutiny of Data</b></td><td>
<ul>
<li>checking accuracy and consistency of data</li>
<li>There is no rule for it, one must apply his intelligence, patience and experience while scrutinizing the given information.</li>
<li>Internal Consistency: When two or more series of related data are given, we should check consistency among them.</li>
</ul>
</td></tr>
<tr><td><b>Presentation of Data – Classification / Organization of Data</b></td><td>
<b>Classification or Organisation</b>: putting data in a neat, precise, and condensed form, making it comparable, suitable for analysis, more understandable.
<table>
<tr><td><b>Chronological/ Temporal/ Time Series Data</b></td><td><ul><li>Data arranged based on Time</li><li>Example: Revenues YoY i.e year on year</li></ul></td></tr>
<tr><td><b>Geographical or Spatial Series Data</b></td><td><ul><li>Arrangement based on regions</li><li>Example: Country wise Revenue of a global company</li></ul></td></tr>
<tr><td><b>Qualitative or Ordinal Data</b></td><td><ul><li>Based on some attribute</li><li>Nationality Wise Medal Winners in Olympics</li></ul></td></tr>
<tr><td><b>Quantitative or Cardinal Data</b></td><td><ul><li>Based on some variable</li><li>Example: Frequency Distribution of a Data</li></ul></td></tr>
</table>
</td></tr>
<tr><td><b>Mode of Presentation of Data – Textual</b></td><td>
<ul>
<li>This method comprises presenting data with the help of a paragraph or several paragraphs.</li>
<li>This is not a suitable mode of presentation as it is dull, monotonous and non-comparable.</li>
</ul>
</td></tr>
<tr><td><b>Mode of Presentation of Data – Tabular Form</b></td><td>
<ul>
<li>When data is shown in the form of <b>Table</b>.</li>
<li>Useful in easy comparison</li>
<li>Complicated data can be presented</li>
<li>Table is must to create a diagram</li>
<li>No analysis possible without table</li>
<li>Components of Table</li>
</ul>
</td></tr>
<tr><td><b>Components of Table</b></td><td>
<table>
<tr><th>Description</th><th>Name of Component of Table</th></tr>
<tr><td><b>Entire Upper Part</b></td><td>Box Head</td></tr>
<tr><td><b>Upper Part describing columns and sub-columns</b></td><td>Caption</td></tr>
<tr><td><b>Left part of the table describing rows</b></td><td>Stub</td></tr>
<tr><td><b>Main Data of Table</b></td><td>Body</td></tr>
<tr><td><b>Source of Data at the bottom of Table</b></td><td>Footnote</td></tr>
</table>
</td></tr>
</table>

Mode of Presentation of Data – Diagrams	• Can be used by educated and uneducated section of society • Hidden trend can be traced • If priority is accuracy, then tabulation is better
Line Diagram	• Time Series is generally in x axis • For wide fluctuation – log chart or ratio chart is used • Two or more series of same unit – Multiple Line Chart • Two or more series of different unit – Multiple Axis Chart
Bar Diagram	• Bar means rectangle of same width and of varying length drawn horizontally or vertically • For comparable series – multiple or grouped bar diagrams can be used • For data divided into multiple components – subdivided or component bar diagrams • For relative comparison to whole, percentage bar diagrams or divided bar diagrams • Vertical Bar Diagram: Useful for Data varying over Time and Quantitative Data • Horizontal Bar Diagram: Useful for Data varying over Space and Qualitative Data
Pie Chart	• Used for circular presentation of relative data (% of whole) • Summation of values of all components/segments are equated to 360 Degree (total angle of circle) • **Segment angle =** $\dfrac{\textbf{(segment value x 360°)}}{\textbf{(total value)}}$

Statistical Description of Data – Frequency Distribution

Frequency and Distribution	• Frequency means number of times a particular observation is repeated. • Frequency Distribution is table which contains observation or class intervals in one column and corresponding frequency in the other. • Definition: A frequency distribution may be defined as a – tabular representation of statistical data, usually in an ascending order, – relating to a measurable characteristic – according to individual value or a group of values of the characteristic under study.

<table>
<tr><td><b>Types of Frequency Distribution</b></td><td>
<table>
<tr><td><b>Ungrouped/ Simple Frequency Distribution</b></td><td>• When there are limited number of distinct observations, frequency can be assigned to each one of them.
• This distribution is simple</td></tr>
<tr><td><b>Grouped Frequency Distribution</b></td><td>• When there are large no. of observations, grouping is done among them (generally in ascending order).
• Each group is called as class interval and frequency is assigned to group and not individual values,
• this is called Grouped Frequency Distribution</td></tr>
</table>
</td></tr>
<tr><td><b>Class Limit</b></td><td>
• For a class interval CL is the minimum and maximum value the class interval may contain
• Minimum Value – Lower Class Limit
• Maximum Value – Upper Class Limit
<table>
<tr><th>Class Interval</th><th>Frequency</th><th>LCL</th><th>UCL</th></tr>
<tr><td>10-19</td><td>10</td><td>10</td><td>19</td></tr>
<tr><td>20-29</td><td>5</td><td>20</td><td>29</td></tr>
<tr><td>30-39</td><td>8</td><td>30</td><td>39</td></tr>
</table>
</td></tr>
<tr><td><b>Classification of Grouped of Frequency Distribution</b></td><td>
<table>
<tr><td>Mutually Exclusive / Overlapping Classification
20 exclude
<table>
<tr><th>Class</th><th>LCL</th><th>UCL</th></tr>
<tr><td>10-20</td><td>10</td><td>20</td></tr>
<tr><td>20-30</td><td>20</td><td>30</td></tr>
<tr><td>30-40</td><td>30</td><td>40</td></tr>
</table>
</td><td>• Here UCL an interval and LCL of next interval are same
• This is usually applicable for continuous variable.
• An observation which is equivalent to common class limit is excluded from the class interval where it is UCL and taken in the class where it is LCL.</td></tr>
<tr><td>Mutually Inclusive / Non-Overlapping Classification
<table>
<tr><th>Class</th><th>LCL</th><th>UCL</th></tr>
<tr><td>10-19</td><td>10</td><td>~~20~~ 19</td></tr>
<tr><td>20-19</td><td>20</td><td>~~30~~ 29</td></tr>
<tr><td>30-39</td><td>30</td><td>~~40~~ 39</td></tr>
</table>
</td><td>• There is no common class limits between two intervals.
• This is usually applicable to discrete variable.
• All observation including UCL and LCL will be taken in the same class interval as there is no confusion.</td></tr>
</table>
</td></tr>
<tr><td><b>Class Boundary</b></td><td>
<table>
<tr><td>In case of Exclusive / Overlapping Classification</td><td>Class Boundary = Class Limit
<table>
<tr><th>Class</th><th>LCL</th><th>UCL</th><th>LCB</th><th>UCB</th></tr>
<tr><td>10-20</td><td>10</td><td>20</td><td>10</td><td>20</td></tr>
<tr><td>20-30</td><td>20</td><td>30</td><td>20</td><td>30</td></tr>
<tr><td>30-40</td><td>30</td><td>40</td><td>30</td><td>40</td></tr>
</table>
</td></tr>
<tr><td>In case of Inclusive / Overlapping Classification</td><td>Lower Class Boundary
LCB = LCL – 0.5
UCB = UCL + 0.5
<table>
<tr><th>Class</th><th>LCL</th><th>UCL</th><th>LCB</th><th>UCB</th></tr>
<tr><td>10-19</td><td>10</td><td>19</td><td>9.5</td><td>19.5</td></tr>
<tr><td>20-29</td><td>20</td><td>29</td><td>19.5</td><td>29.5</td></tr>
<tr><td>30-39</td><td>30</td><td>39</td><td>29.5</td><td>39.5</td></tr>
</table>
</td></tr>
</table>
</td></tr>
</table>

<table>
<tr><td><b>Mid-Point / Class Mark / Mid Value of Class Interval</b></td><td>

<table><tr><td>$\frac{\text{LCL+UCL}}{2}$</td><td>$\frac{\text{LCB+UCB}}{2}$</td></tr></table>

<ul><li>Useful in calculation of AM, GM, HM, SD in case of grouped frequency distribution</li></ul></td></tr>
<tr><td><b>Class Length/ Width or Size</b></td><td><b>UCB – LCB only</b></td></tr>
<tr><td><b>Cumulative Frequency</b></td><td>
<ul><li>Less than type: It shows no. of observations less than UCB</li>
<li>More than type: It shows no. of observations more than UCB</li></ul>

<table>
<tr><th>Class Interval</th><th>Freq.</th><th>UCB</th><th>Less than type CF</th><th>More than type CF</th><th>Total of both CF</th></tr>
<tr><td>44-48</td><td>3</td><td>48.5</td><td>3</td><td>33</td><td>36</td></tr>
<tr><td>49-53</td><td>4</td><td>53.5</td><td>7</td><td>29</td><td>36</td></tr>
<tr><td>54-58</td><td>5</td><td>58.5</td><td>12</td><td>24</td><td>36</td></tr>
<tr><td>59-63</td><td>7</td><td>63.5</td><td>19</td><td>17</td><td>36</td></tr>
<tr><td>64-68</td><td>9</td><td>68.5</td><td>28</td><td>8</td><td>36</td></tr>
<tr><td>69-73</td><td>8</td><td>73.5</td><td>36</td><td>0</td><td>36</td></tr>
<tr><td>Total</td><td>36</td><td></td><td></td><td></td><td></td></tr>
</table></td></tr>
<tr><td><b>Frequency Density</b></td><td>$\frac{\text{Class Frequency}}{\text{Class Length of class}}$</td></tr>
<tr><td><b>Relative Frequency</b></td><td>$\frac{\text{Class frequency}}{\text{Total Frequency}}$
Its can have values between 0 and 1</td></tr>
<tr><td><b>Percentage Frequency</b></td><td>$\frac{\text{Class frequency}}{\text{Total Frequency}} \times 100$</td></tr>
<tr><td><b>Frequency Dist. Diagram – Histogram</b> /area diag</td><td><ul><li>It is a convenient way to represent FD</li><li>Comparison between frequency of two different classes possible</li><li>It is useful to calculate mode also</li></ul></td></tr>
<tr><td><b>Frequency Polygon</b></td><td><ul><li>Usually preferable for ungrouped frequency distribution</li><li>Can be used for grouped also but only if class lengths are even</li></ul></td></tr>
<tr><td><b>Ogives/ Cumulative Frequency</b></td><td><ul><li>This graph can be made by both type of Cumulative Frequency and called as Less than Ogive or More than Ogive</li><li>It can be used for calculating quartiles, median</li></ul></td></tr>
<tr><td><b>Frequency Curve</b></td><td><ul><li>It is a limiting form of Area Diagram (Histogram) or Frequency Polygon</li><li>It is obtained by drawing smooth and free hand curve though the mid points</li><li>Most used curve is Bell Shaped</li></ul></td></tr>
</table>

Index Numbers

Practical Examples of Index Numbers	• Index numbers are convenient devices for **measuring relative changes (generally in %)** of differences from **time to time** or from **place to place** • Series of numerical figures which show relative position • Index Numbers show percentage changes rather than absolute amounts of change
Data Selection	• It **depends on the purpose** for which the index is used. • Index numbers are often constructed from the **sample**. **Random sampling,** and if need be, a **stratified random sampling** can be used to ensure that sample is representative. • Data should be **comparable** by ensuring consistency in selection method.
Base Period	• It is a **point of reference** in comparing various data. • Standard point of comparison. • The period should be **normal**. • It should be **relatively recent** • Choice of suitable base period is a temporary solution
Use of Averages	• The **geometric mean is better** in averaging relatives, • But for most of the index's **arithmetic mean is used because of its simplicity**
Price/ Quantity/ Value Relative	For Individual Commodity, $\dfrac{\textbf{Current Period Price/ Quantity/ Value}}{\textbf{Base Period Price/ Quantity/ Value}}$
Link Relative	$\frac{P_1}{P_0}, \frac{P_2}{P_1}, \frac{P_3}{P_2}, \dots, \frac{P_n}{P_{n-1}}$ Same can be created for quantities also
Chain relatives	When the above relatives are in respect to a fixed base period these are also called the chain relatives $\frac{P_1}{P_0}, \frac{P_2}{P_0}, \frac{P_3}{P_0}, \dots, \frac{P_n}{P_0}$
Formula for Chain Index (when direct data is not available)	$\dfrac{\textbf{Link relative of current year × Chain Index of previous year}}{\textbf{100}}$ The chain index is an unnecessary complication unless of course where data for the whole period are not available or where commodity basket or the weights have to be changed.
Limitations of Index Numbers	• Chances of errors due to Sampling • It gives broad trend not real picture • Due to many methods, at times it creates confusion
Usefulness of Index Numbers	• Index numbers are very useful in deflating (eg. Nominal wages into real) • Framing suitable policies in economics and business • They reveal trends and tendencies in making important conclusions

	• They are used in time series analysis to study long-term trend, seasonal variations and cyclical developments
Formula for Deflated Value	$\textbf{Deflated Value} = \dfrac{\textbf{Current Value}}{\textbf{Price Index of the current year}}$
Shifted Price Index	$\dfrac{\textbf{Original Price Index}}{\textbf{Price Index of the year on which it has to be shifted}} \times \textbf{100}$
Unit Test	• This test requires that the formula should be independent of the unit in which or for which prices and quantities are quoted. • Except for the simple (unweighted) aggregative index all other formulae satisfy this test.
Time Reversal Test	• It is a test to determine whether a given method will work both ways in time, forward and backward. • $P_{01} \times P_{10} = 1$ • Laspeyres' method and Paasche's method do not satisfy this test, but Fisher's Ideal Formula does.
Factor Reversal Test	• This holds when the product of price index and the quantity index should be equal to the corresponding value index. • Symbolically $\mathbf{P_{01} \times Q_{01} = V_{01}}$ • Fisher's Index Number is ideal as it satisfies Unit, Time Reversal and Factor Reversal Test
Circular Test	• This property therefore enables us to adjust the index values from period to period without referring each time to the original base. • It is an extension of time reversal test • The test of this **shiftability of base** is called the circular test. • This test is not met by Laspeyres, or Paasche's or the Fisher's ideal index. • The weighted GM of relative, **simple geometric mean of price relatives** and the **weighted aggregative with fixed weights meet this test.** (These methods are not in syllabus)
Cost of Living Index (also called General Index)	• CLI is defined as the **weighted AM of index numbers** of few groups of basic necessities. • AM of group indices gives the General Index • Generally, for calculating CLI; food, clothing, house rent, fuel & lightning and miscellaneous groups are taken into consideration. • Examples of CLI: WPI, CPI, etc.
Symbol	• P_{01} is the index for time 1 on 0 • P_{10} is the index for time 0 on 1

P_{01}

Measures of Central Tendency

Arithmetic Mean

Property 1	If all the observations are constant, AM is also constant
Property 2	the algebraic sum of deviations of a set of observations from their AM is zero
Property 3	AM is affected both due to change of origin and scale If $y=a+bx$ then $\bar{y}=a+b\bar{x}$
Property 4	Combined AM $\bar{x}_c=\dfrac{n_1\bar{x}_1+n_2\bar{x}_2}{n_1+n_2}$
General Review	• AM is best measure of central tendency • AM is based on all observations • AM is affected by sampling fluctuations • AM is amenable to mathematical property • AM cannot be used in case of open end classification

Median

Property 1	For a set of observations, the sum of absolute deviations is minimum, when the deviations are taken from the median. $\sum\lvert x_i-Me\rvert$ modulus
Property 2	Median is also affected by both change of origin and scale.
General Review	• Median is also called as positional average • Median is not based on all observations • Median is not affected by sampling fluctuations • Median is best measure of central tendency in case of open end classification

Partition Values

<table>
<tr><td><b>Meaning</b></td><td>
<ul>
<li>These may be defined as <b>values dividing</b> a given <b>set of observations</b> into number of <b>equal parts</b></li>
<li>When we want to divide the given set of observations into two equal parts, we consider median, similarly there are quartiles, deciles, percentiles</li>
</ul>
<table>
<tr><th>Name of PV</th><th>No. of equal parts</th><th>No. of PVs</th><th>Symbol</th></tr>
<tr><td>Median</td><td>2</td><td>1</td><td><b>Me</b></td></tr>
<tr><td>Quartile</td><td>4</td><td>3</td><td>Q_1, Q_2, Q_3</td></tr>
</table>
</td></tr>
</table>

		Decile	10	9	$D_1, D_2, \ldots, D_9$
		Percentile	100	99	$P_1, P_2, \ldots, P_{99}$

Mode – Concept/ Formula

Meaning	Mode is the **value** that **occurs the maximum** number of times
Special Thing about Mode	• If two or more observations are having maximum frequency then there are **multiple modes** [multimodal distribution] • If there are **exactly two** modes then distribution is called as **Bimodal** Distribution • If all observations are having same frequency then distribution has **no mode** • We can say that Mode is **not rigidly defined**
Property 1	If all the observations are constant, mode is also constant
Property 2	Mode is also affected both due to change of origin and scale
General Review	• Mode is not based on all observations • Mode is not rigidly defined • Mode is not amenable to Mathematical Property

Relationship between Mean, Median and Mode

In case of Symmetric Distribution	**Mean = Median = Mode**
In case of Moderately Skewed Distribution (Empirical relationship)	**Mean – Mode = 3 (Mean – Median)**

Geometric Mean

Definition	For a given set of n **positive observations**, the geometric mean is defined as the n^{th} root of the product of the observations
Property 1	Logarithm of G for a set of observations is the AM of the logarithm of the observations $\log G = \frac{1}{n}\sum \log x$
Property 2	If all the observations are constant, GM is also constant
Property 3	$\text{GM of } z = \text{GM of } x \times \text{GM of } y$
Property 4	$\text{GM of } z = \frac{\text{GM of } x}{\text{GM of } y}$

Harmonic Mean

Definition	For a given set of **non-zero** observations, harmonic mean is defined as the **reciprocal of the AM of the reciprocals of the observation**
Property 1	If all observations are constant HM is also constant

Use of GM and HM

Both	Both are used for calculating average rates
GM	Appropriate for rates having percentages
HM	Appropriate for rates other than percentages

Measures of Dispersion

Meaning of Measure of Dispersion	- Dispersion for a given set of observations may be defined as - the **amount of deviation** of the observations, - usually, from an **appropriate** measure of **central tendency**	
Types of Measure of Dispersion	**Absolute Measures of Dispersion**	- These are with units - These are not useful for comparison of two variables with different units. - Example: Range, Mean Deviation, Standard Deviation, Quartile Deviation
	Relative Measures of Dispersion	- These are unit free measures - These are useful for comparison of two variables with different units. - Example: Coefficient of Range, Coefficient of Mean Deviation, Coefficient of variation, Coefficient of Quartile Deviation

Range

Property 1	- **Not affected** by change of **origin** - Affected by **change of scale (only value)** - **No impact of sign** of change of scale - Note: **Measure of Dispersion can never be negative**
General Review	- Not Based on All Observations - Easy to Compute

Mean Deviation

Meaning	- Mean deviation is defined as the - **arithmetic mean** of the - **absolute deviations** of the observations - from an **appropriate measure** of central tendency
Property 1	Mean Deviation takes its **minimum value** when deviations are taken from **Median**

Property 2	Change of Origin – **No Affect**, Change of Scale – **Affect of value not sign**
General Review	• Based on **all observations** • Improvement over Range • **Difficult to compute** • **Not amenable to Mathematical Property** because of usage of **Modulus**

Standard Deviation

Meaning	• Improvement over Mean Deviation • It is defined as the **root mean square deviation** when the deviations are taken from the AM of the observations
Coefficient of Variation	$\frac{SD_x}{\bar{x}} \times 100$
SD for any two numbers	$SD = \frac{\lvert a-b \rvert}{2}$
SD for first n natural numbers	$s = \sqrt{\frac{n^2-1}{12}}$
Property 1	If all the observations are constant, SD is **ZERO**
Property 2	No effect of change of origin but affected by change of scale in the magnitude (ignore sign)
Property 3	$SD_c = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2 + n_1 d_1^2 + n_2 d_2^2}{n_1 + n_2}}$ $d_1 = \bar{x}_c - \bar{x}_1$ $d_2 = \bar{x}_c - \bar{x}_2$

Quartile Deviation

Meaning	It is semi-inter quartile range
General Review	• It is the **best measure** of dispersion for **open-end** classification • It is also **less affected** due to sampling fluctuations • Like other measures of Dispersion, **QD** is also not affected by change of origin but affected by scale ignoring sign

Correlation and Regression

Bivariate Data

Definition	- When data are collected on two variables **simultaneously**, they are known as **bivariate data** - and the corresponding frequency distribution, derived from it, is known as **Bivariate Frequency Distribution**
Marginal Distribution	- It is the frequency distribution of **one variable** (x or y) across the other variable's **full range of values** - **Number of Marginal Distribution = 2**
Conditional Distribution	- It is the frequency distribution of **one variable** (x or y) across a **particular sub-population** of the other variable. - **No. of Conditional Distributions = m + n** ***m = no. of class interval of x*** ***n = no. of class interval of y***

Scatter Diagram

Concept Points	- It helps us to find **Nature** and **Relative Strength** of Correlation - It is useful for **Non-Linear** Correlation also - It **cannot** be used to determine **value** - Diagrams are **time taking**

Karl Pearson's Correlation Coefficient

<table>
<tr><td><b>How to Calculate</b></td><td>Correlation Coefficient is the ratio of covariance with product of standard deviations</td></tr>
<tr><td><b>Property 1</b></td><td>The Coefficient of Correlation is a <b>unit-free measure</b></td></tr>
<tr><td><b>Property 2</b></td><td>Value lies from <b>-1 to +1</b></td></tr>
<tr><td><b>Property 3</b></td><td><table>
<tr><td><b>Change of Origin</b></td><td>No impact</td></tr>
<tr><td><b>Change of Scale</b></td><td>No impact of value, but if change of scale of both variables are of <b>different sign</b> then <b>sign of r</b> will also change</td></tr>
</table></td></tr>
<tr><td><b>Interpretation of Value of r</b></td><td><table>
<tr><th>Value of r</th><th>Interpretation</th></tr>
<tr><td>-1</td><td>Perfect Negative</td></tr>
<tr><td>Between -1 and 0</td><td>Negative</td></tr>
<tr><td>Closer to -1</td><td>Strong Negative</td></tr>
<tr><td>Far from -1</td><td>Weak Negative</td></tr>
<tr><td>0</td><td>No Correlation</td></tr>
<tr><td>Between 0 and 1</td><td>Positive</td></tr>
<tr><td>Far from +1</td><td>Weak Positive</td></tr>
<tr><td>Near to +1</td><td>Strong Positive</td></tr>
<tr><td>+1</td><td>Perfect Positive</td></tr>
</table></td></tr>
</table>

Spearman's Rank Correlation Coefficient

Usage	• find the level of **agreement (or disagreement**) between two judges so far as assessing a **qualitative characteristic (attribute)** is concerned • Use in case of ranks
Ranking in case of Tie	In case of tie, simple average of ranking should be assigned to tied values

Coefficient of Concurrent Deviations

Usage	A very **quick, simple** and **casual** method of finding correlation when we are not serious about the magnitude of the two variables

Regression Basics

<table>
<tr><td><b>Meaning</b></td><td>Estimation of one variable for a <b>given value</b> of another variable on the basis of an <b>average mathematical relationship</b> between the two variables</td></tr>
<tr><td><b>Requirements</b></td><td>• Estimation of Y when X is given
• Estimation of X when Y is given</td></tr>
<tr><td><b>General Points</b></td><td><table>
<tr><td><b>Perfect Correlation</b></td><td>• When linear relationship exists between two variables, correlation is perfect.
• Perfect Correlation is represented by a linear equation and this equation can be used for regression purpose directly.
• Same equation can be used in both ways</td></tr>
<tr><td><b>Imperfect Correlation</b></td><td>• In case of imperfect correlation there is no definite line and equation
• We will use method of least square to estimate both regression lines</td></tr>
</table></td></tr>
<tr><td><b>Formula of Regression Equations/ Lines</b></td><td><table>
<tr><td>Estimation of Y when X is given</td><td>• Use Regression line of <b>Y on X</b>
• Equation Format:
$Y-\bar{Y}=b_{yx}(X-\bar{X})$
b_{yx} is regression coefficient of Y on X</td></tr>
<tr><td>Estimation of X when Y is given</td><td>• Use Regression line of <b>X on Y</b>
• Equation Format:
$X-\bar{X}=b_{xy}(Y-\bar{Y})$
b_{xy} is regression coefficient of X on Y</td></tr>
</table></td></tr>
<tr><td><b>Property 1</b></td><td>Change of Origin and Scale
• Origin: No Impact
• Scale: If original pair is $x,\ y$ and modified pair is u, v
$b_{vu}=b_{yx}\times\frac{\text{change of scale of y}}{\text{change of scale of x}}$</td></tr>
</table>

	$b_{uv} = b_{xy} \times \frac{\text{change of scale of x}}{\text{change of scale of y}}$
Property 2	Two regression lines (if not identical) will intersect at the point [means] $(\bar{x}, \bar{y})$
Property 3	Relation between Correlation and Regression Coefficients $r_{xy} = \pm\sqrt{b_{xy} \times b_{yx}}$ r_{xy}, b_{xy}, b_{yx} will always have same sign

Probable Error

<table>
<tr><td><b>Use</b></td><td><ul><li>Correlation is calculated using sample, value for sample may differ from population, this difference is probable error</li><li>If there is significant probable error, there is no evidence of real correlation</li></ul></td></tr>
<tr><td><b>Limits of Sample Correlation Coefficient</b></td><td>$r \pm PE$</td></tr>
<tr><td><b>How to check evidence of Correlation using PE</b></td><td><table>
<tr><th>Case</th><th>Conclusion</th></tr>
<tr><td>If r is less than PE</td><td>There is no evidence of correlation</td></tr>
<tr><td>If r is greater than six times of PE</td><td>The presence of correlation is certain</td></tr>
<tr><td>Since r lies from -1 to +1</td><td>PE can never be negative</td></tr>
</table></td></tr>
</table>

Coefficient of Determination and Non-Determination

Coefficient of Determination Accounted Variance/ Explained Variance	r^2
Coefficient of Non-Determination Unaccounted Variance/ Unexplained Variance	$1 - r^2$